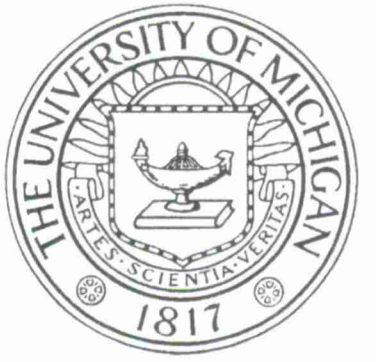
Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components

Final Report
Project N00014-10-1-0152

David E. Kieras
University of Michigan



Report No. FR-12/ONR-EPIC-18

Period Covered: 1 NOV 2009 – 30 SEP 2012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	20 March 2013	Final Report 1 Nov 2009 – 30 Sep 2012

4. TITLE AND SUBTITLE
Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components

5. FUNDING NUMBERS
N00014-10-1-0152

10PR02787-00

6. AUTHOR(S)
David E. Kieras

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan Division of Research Development and Administration, Ann Arbor, MI 48109

8. PERFORMING ORGANIZATION REPORT NUMBER
FR-12/ONR-18

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research (Code 341)
875 N. Randolph St.
Arlington, VA 22203-1995

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This is the final report for a project that was in a series of projects on the development and validation of the EPIC cognitive architecture for modeling human cognition and performance. This project focussed on extending the architecture to account for sound and speech phenomena, with emphasis on multichannel speech comprehension in a simple command-and-control task for which considerable empirical data is available. Additional work concerned application of the EPIC architecture to Navy research problems.

14. SUBJECT TERMS
Cognitive Architecture, Human Performance Modeling

15. NUMBER OF PAGES
34

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components
Final Report

ONR Grant N00014-10-1-0152
Period Covered: 1 November 2009 – 30 September 2012

David Kieras, Principal Investigator
Gregory H. Wakefield, Co-Principal Investigator

Introduction

This is the final report for a project on the development and validation of the EPIC cognitive architecture for modeling human cognition and performance. It continued a series of ONR-sponsored projects on the development of the EPIC architecture for human cognition and performance, begun by the same PI with Co-PI David Meyer of the University of Michigan; this project included Gregory Wakefield of the University of Michigan as Co-PI. This extended activity produced a large number of products and accomplishments; however, this report documents the outcome of only this specific project.

The general goal of this project was to extend the EPIC computational architecture for modeling human cognition and performance and apply it to tasks relevant to military applications of computational cognitive modeling. The highest-priority specific goal was to develop the EPIC architecture so that it supports predictive modeling of human-computer interaction involving spatialized audio and speech activity and does so at least as well as it currently supports modeling the visual, manual, and procedural aspects of such tasks. Secondary goals were in the closely related areas of working memory in task performance and spatial aspects of manual motor activity.

Administrative Note. The plan was to continue this specific project on the same basic topics, but due to various delays, the status of the new project was unclear for long enough that rather than continue the prior project administratively, it was considered terminated and a new project was approved and funded effective 1 January 2013. Not only did this process involve a hiatus in the project activity, but also the delayed the decision that the prior project was not going to be continued, resulting in a delay in this final report. The new project, N00014-13-1-0358, will be continuing the lines of research described here.

Concise Summary of Project Accomplishments

This section provides a brief summary of what was accomplished under each goal of the original proposal. The next section provides an extended treatment of the major work accomplished.

Goal 1. *The EPIC architecture components for audition and speech communication will be expanded to provide a robust modeling and prediction capability for tasks involving speech production and speech and non-speech audio input with special attention to the role of spatialized audio.*

Considerable progress was made on architectural models of multichannel speech comprehension, but the effort proved to be difficult enough that it must be continued in a new project, and work on other aspects of this goal was postponed, in particular spatial location in multichannel speech processing. A second major topic under this goal, modeling simultaneous speaking and listening, proved to have weak empirical underpinnings, leading to a suspension of the modeling work and the launching of collaborative efforts with researchers Nandini Iyer and Brian Simpson at Wright-Patterson AFB to explore this issue empirically.

Goal 2. *As needed to account for the speech-related phenomena, verbal working memory and task working memory mechanisms will be developed and added to the architecture.*

This goal was included as a second priority in the original proposal, anticipating that it would be needed for the speaking-while-listening task, and the application to a complex CIC-like task in Goal 4. It was suspended when it became clear that the speaking-while-listening phenomena needed further empirical work, and the dataset for the complex CIC-like task was not yet available.

Goal 3. *The architectural mechanism underlying spatially-based manual motion will be further developed and explored to account for aimed movement efficiency, spatial coding of responses such as button-pushing, and spatial S-R compatibility, and the relationship of these spatial factors with the spatial representations underlying visual and auditory localization.*

This goal was included as a second priority in the original proposal, and since the Goal 1 work on auditory localization was postponed, no additional work was pursued on these topics.

Goal 4. *The models and architecture will be developed and tested in complex military-like tasks, e.g. based on Combat Information Center (CIC) Anti-Air Warfare (AAW) tasks, based on earlier work with the Multi-Modal Watch Station (MMWS) project.*

Some modeling work involving localized sound integrated with visual search was performed on this goal, primarily with Dr. Michael Qin's group at NSMRL, with some discussion and planning of a larger-scale CIC-like task with Dr. Derek Brock of NRL.

Goal 5. *The EPIC architecture software and models will be updated, improved, and made easily available as needed to support this work and that of other researchers. Similar work will be done as appropriate for the related practically-oriented GLEAN software.*

As expected, only a small effort was required during the project period; substantial updating will be involved when the auditory architecture has stabilized. Some work was done to move the software into the new C++11

Standard, and for the Mac OS X implementation, preparations made to move to the newer Cocoa API instead of the previous Carbon API.

In the following sections, an extended treatment will be presented only for each goal in which significant progress was made, namely Goals 1 and 4.

Extended Summary of Major Accomplishments

Goal 1. The EPIC architecture components for audition and speech communication will be expanded to provide a robust modeling and prediction capability for tasks involving speech production and speech and non-speech audio input with special attention to the role of spatialized audio.

Approach

Prior to this project, both EPIC and the closely related GLEAN architecture (Kieras, Wood, Abotel, & Hornof, 1995; Kieras & Knudsen, 2006) had basic facilities for presenting simulated audio signals and speech input to the simulated human, and generating simulated speech output from the human which can be sent to the simulated task environment or other simulated humans. In various models developed over many years, these architecture facilities demonstrated their value, such as simulated telephone operators in Kieras, Wood, and Meyer (1997), and CIC team members interacting via speech in Santoro, Kieras, and Pharmer (2004). However, the very simple form of these facilities did not address many critical issues that the future design of military human-computer systems will have to deal with. The basic problem is that the important phenomena specific to audition, such as spatial localization, masking, and segregation of auditory streams of both speech and non-speech sound were not represented. These phenomena are astoundingly complex and subtle, but nonetheless, must be represented in a useful cognitive architecture because current and future military task settings, such as CIC watch standing, involve multiple simultaneous speech inputs.

In this project, we made good progress in combining the two relevant scientific bases for developing an auditory cognitive architecture. Kieras contributed expertise in extending the cognitive architecture for the auditory and speech mechanisms and constructing the models of human performance in the experimental tasks, while Wakefield, an audition and mathematical modeling expert, brought an understanding of the complex phenomena and signal-processing analysis underlying the conventional audition and speech comprehension research. Working together, we developed ways to combine the methodology of mathematical signal analysis of perception with the cognitive architecture and task strategy mechanisms to produce an initial architecture and model for an important 2-speaker task whose scope and quantitative accuracy is unprecedented in the audition and speech science field. A summary of this progress is presented below.

Overview of the EPIC Architecture

The EPIC architecture is summarized in Figure 1. A simulated human consisting of several processors is on the right, and a simulated task environment (often called *the device*) is on the left. The EPIC software runs all of the processors, including the device, in simulated parallel as a discrete-event simulation. Each box in the diagram corresponds to a component or set of components in the software that simulates the activity of a subsystem of the simulated human. EPIC is a *performance* modeling system whose goal is to provide a comprehensive account of

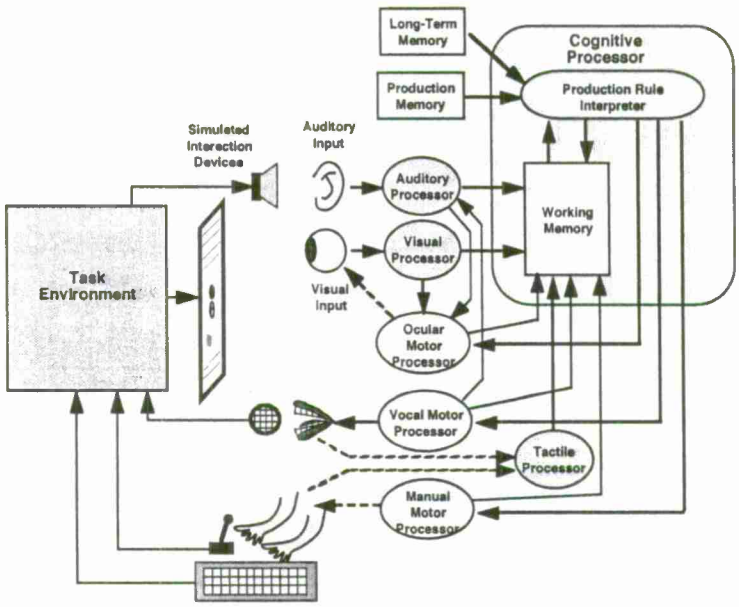


Figure 1. The overall structure of the EPIC architecture. A simulated task environment - usually a device - is on the left; the simulated human, made up of a set of components, is on the right. All components of the simulated human and the device run in parallel. Note the connections between the Vocal Motor Processor, the Auditory Processor, and the Ocular Motor Processor.

human abilities and limitations in perception, cognition, and action; by setting aside the usual preoccupation with learning, it has been possible to develop the architecture to provide useful approximations for a wide variety of mechanisms and processes that are important in realistic task settings. More complete discussions of EPIC and its approach are available elsewhere (Kieras & Meyer, 1997a, b; Meyer & Kieras, 1999; Kieras, 2007, in press). The remainder of this discussion will focus on those portions of EPIC relevant to the project work.

Basic Architecture for Speech and Audition

In the original (LISP) version of EPIC, the components for audition and speech were adequate for modeling some tasks involving auditory signals and speech interaction (e.g. Kieras, Wood, & Meyer, 1997) but were not particularly well-structured. These components were further complicated by additions made to incorporate the phonological loop model of verbal working memory (Kieras, Meyer, Mueller, & Seymour, 1999) in which the vocal motor processor deposits encodings of spoken words (either overt or covert) into the auditory working memory, as shown by the connection in Figure 1. Further modifications were made by Mueller (2002) to support the complex retrieval and guessing strategies required to fit detailed patterns of recall in verbal working memory.

Another set of additions handled localized sound; this was done in a project with James Ballas of NRL whose dual-task paradigm has been extensively modeled in EPIC (Kieras & Meyer, 1995; Ballas, Kieras, & Meyer, 1996) and other architectures. This task involved two displays; the subject tracks a moving target in the right-hand display with a right-hand joystick while responding to events on a radar-like display on the left with the left hand. The sound augmentation used spatialized audio cues to designate events as they happened in the radar display, which lead to better dual-task performance that was accounted for with a model in EPIC with localized sound incorporated (Ballas, Brock, Stroup, Kieras & Meyer, 1999; Kieras, Ballas, & Meyer, 2001). This "Ballas task" has been used for

subsequent research on spatialized audio by Anthony Hornof at Oregon (Hornof, Zhang and Halverson, 2010; Hornof and Zhang, 2010) and by Brock's group at NRL (Brock, Ballas, Stroup, & McClimens, 2004). The major addition to the architecture was to combine visual and auditory spatial representations by postulating that an object with a perceived location in space could have both visual and auditory properties. A connection was added between the auditory processor and the oculomotor processor whereby a sound onset could trigger a reflexive eye movement to the location of the sound, and a production rule could specify an eye movement to a sound's location as well; this relationship between the auditory processor and the oculomotor processor is shown in Figure 1.

Figure 2 shows the current tentative internal structure of the auditory system in EPIC. Figure 2 is based largely on analogy with the visual system, but we have determined that the audition and speech literature is in fact rather vague about the levels and stages of processing involved in the auditory system. For example, energetic masking is assumed to happen at the strictly sensory acoustic level, while informational masking is supposed to be more cognitive/perceptual. But since the theoretical discussion lacks an architectural basis, it is difficult to be clear and which parts of the system do what. In fact, in most of our modeling work, we have chosen the first stage (the Ear Processor) as the locus of modifications simply because it is easier to modify only a single module while exploring a variety of recognition functions and stream tracking mechanisms. Future work should clarify what the useful processing stages should be.

Modeling 2-channel speech processing

In our initial work, we surveyed key empirical studies on basic multiple-channel speech processing. This led to a decision that our first model should be of a basic 2-channel speech processing task, whereupon we realized that most of the studies in the literature omitted some key empirical facts concerned the "fate" of the "unattended" material.

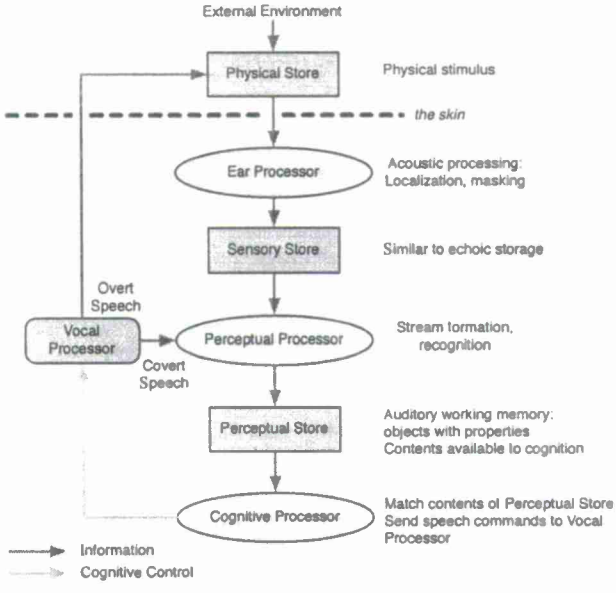


Figure 2. A more detailed view of the auditory and speech components of the EPIC architecture. The processors and stores create the stored representations. Cognition examines the final representation and sends commands to the motor processors. Note how the vocal motor processor sends information directly to auditory perception and the physical environment.

However, one of Brungart's earliest studies (Brungart, 2001) provided an ideal first study to model. This study compared performance in a two-speaker task using the well-known CRM corpus, in which each message consists of sentences of the form *Ready <callsign> go to <color> <digit> now*. The task is to follow the message with the designated target callsign (*baron*) and click on the corresponding colored digit on a display. The two messages differed in relative loudness (signal-to-noise ratio, SNR) and three levels of speaker similarity: Different Sex (DS), different speakers of the Same Sex (SS), and the Same Talker (ST). Figure 4 below shows this data as the observed curves.

Unlike almost all studies of multiple speaker processing, this study included some detailed data about the incorrect responses; the results show not only the proportion of correct responses as a function of speaker similarity and SNR, but also a separate breakout for the color and digits, and whether they were correct (from the target message), or from the masker message, or neither one. A key empirical fact is that the incorrect responses were almost always from the masker message, which places a basic constraint on the cognitive-architectural processes in a model. Another key empirical fact is that the color words were recognized rather differently than the digit words, a result that neither Brungart (2001) nor subsequent researchers had attempted to explain.

A final important and puzzling feature of the data is that in some conditions, performance actually improves with decreasing SNR; for example, if the two messages are from the Same Talker, the target digit identification performance at -12 SNR is substantially better than at 0 SNR. Apparently, a message that is very quiet compared to a distractor can still be segregated and responded to. This effect stands out for digits in the Same-Talker condition, but note that a milder form of the effect shows in the color responses and in the Same-Sex condition as a flattening of the accuracy curve with negative SNR. The more usual effect is that accuracy would continue to fall towards zero in an ogival curve with decreasing SNR; in fact, this is the effect produced if the masker for a CRM target is simply noise (Brungart, 2001). Thus the presence of a second masking CRM message produces a different, and puzzling, improvement in the accuracy of identifying the target content.

We focused on trying to capture the detailed results in a model that has been presented at program reviews and invited talks. At the top level, the significance of the model is that it is apparently the *first* effort to marry the type of models typically used in audition and speech work (essentially mathematical psychophysical models based on sound characteristics) with the type of cognitive architectural models like EPIC, ACT-R, and Soar, in which the cognitive processor implements a task strategy described with production rules. Applying an earlier lesson from EPIC (Meyer & Kieras, 1999; Kieras & Meyer, 2000), even simple tasks can have sophisticated strategies, whose qualitative or logical nature is difficult to capture in conventional mathematical models. In summary, we built a psychophysical front end embedded in a set of information-processing stages, with a cognitive-strategic back end, to produce a model that performed the entire task end-to-end. To make this a bit more concrete, Figure 3 shows a screenshot of the EPIC model execution displays for a Brungart task execution.

The basic hypothesis represented in the model is that speech processing requires that two recognitions be done for each word of a speech message: the *content* of the word itself (i.e. recognizing that a sound was the word green), and the *stream ID* that the word was from (which requires discriminating which speaker uttered the word). Each of these recognitions was assumed to depend on the SNR and the speaker conditions (e.g. Same Talker vs Different Sex), and was assumed to be independent of the other recognition and of the recognitions for previous words.

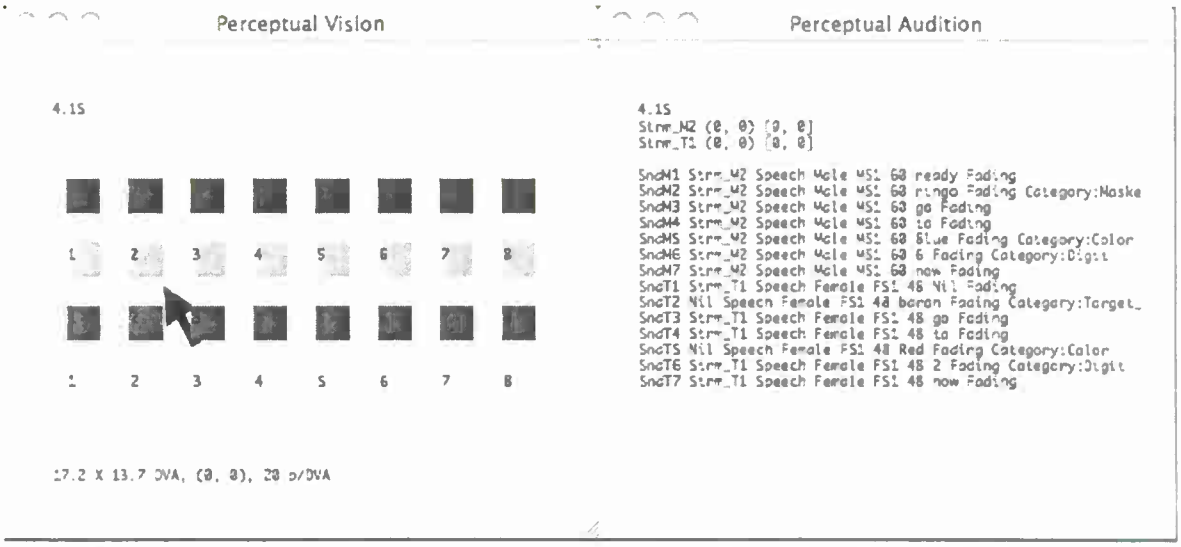


Figure 3. A screenshot of the EPIC model execution windows for the model of the Brungart task. The visual display appears on the left; the current contents of the auditory perceptual store are shown on the right; auditory objects appear for both the target stream and the masker stream. The production rules must try to assign each word to the correct stream even when some of the word content or the stream identifying information is missing.

During the presentation of the messages, auditory memory would fill with a mélange of representations of words with/without recognized content and with/without recognized stream membership. The task strategy implemented by the cognitive processor would then sort through the available information to choose what response to make. Because there are only two possible streams involved, the strategy can make strong inferences to compensate for missing information. For example, if one has not heard the callsign for the target stream, but has for the masker stream, then one can infer that if a different stream is heard, then it must be the target stream. At response time, if no streams have been identified as target or masker, a choice of color and digit from the same stream is a better guess than a color and digit from different streams.

To fit this model to the data, we had to choose the form of psychometric function for the content and stream recognition, estimate the parameters for these, and also we also had to experiment with different task strategies for making inferences and choosing responses. Despite the apparent degrees of freedom in the model, it proved to be very difficult to fit the data; previous reports summarize the novel methodology we developed to solve this problem by using mathematical optimization and cognitive architecture simulations in tandem. This allowed us to explore a complex space of task strategies and perceptual functions efficiently, and we arrived at a model for the Brungart (2001) data that fits remarkably well, capturing both the qualitative trends in the data and the exact numerical values, as shown in Figure 4 below.

The perceptual detection functions, shown in Figure 5, are familiar simple forms (e.g. an ogive relating SNR to probability of recognition, with a probability floor) and are well-behaved across the speaker and SNR manipulations, being strictly ordered from Difference Sex to Same Talker, and show improved detection of stream identification through the course of the utterance, which agrees with a common interpretation of stream formation phenomena.

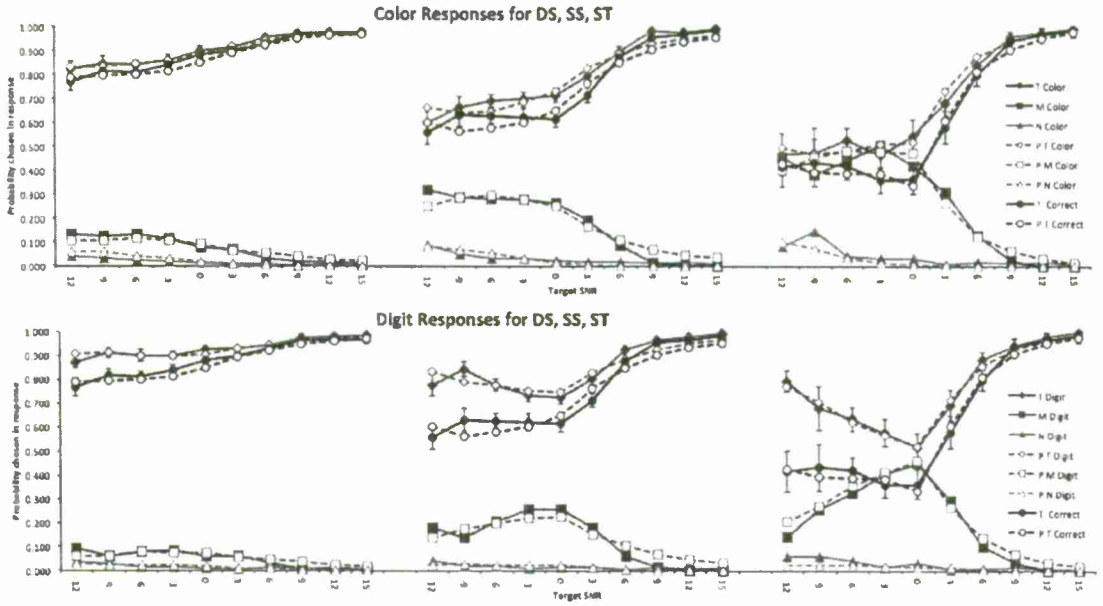


Figure 4. Observed (solid lines and points) and predicted (dotted lines and open points) probabilities for each response choice in the Brungart (2001) data. The vertical axis is probability of response; the horizontal axis is SNR in dB. Color responses are in the upper panel; digit responses are in the lower panel. The three speakers conditions are shown left-to-right in the three subpanels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). Blue curves are for the target color or digit, red for masker color or digit, green for responses that are neither target or masker, and black curves show the probability that both the color and digit are correct (same in both upper and lower panels).

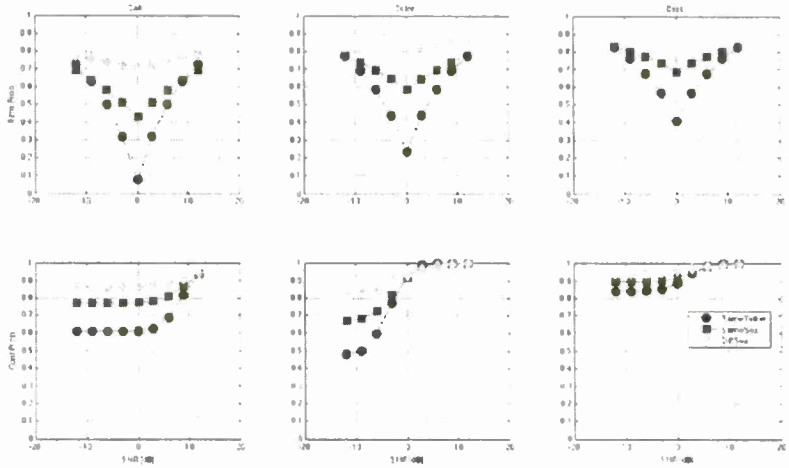


Figure 5. Psychophysical functions used in the model to fit the Brungart (2001) data. The horizontal axis is the SNR in dB; the vertical axis is the probability of recognition. The speaker conditions are color-coded as green, red, and blue for Different Sex, Same Sex, and Same-Talker respectively. The upper panels show the probability of identifying the stream correctly for the callsign word, the color word, and the digit word, from left to right. These functions are an exponential probability function with a minimum probability symmetrical around an SNR of zero. The exponential parameter was constrained to the same value for the three words in each speaker condition, so only the minimum probability varied as a function of word and condition. The lower parameters show the probability of identifying the word content correct for the three words. These are Gaussian functions with a minimum probability whose variance parameter was constrained to the same value for the three words in each condition, so only the mean and minimum probability varied as a function of word and condition.

In summary, the model for 2-channel listening combines psychoacoustic components for perception with cognitive components to implement the inference and decision procedures involved in the task, and provides an excellent account of the data.

Problems with the model

This model looked very promising, but there were some problems. First, it did not in fact contain any mechanism for how streams were identified or "build up" over time, which has long been a concern in the audition field. Rather, we had "finessed" this process with a black-box treatment. Second, the black-boxing of stream ID detection has a couple of odd features in the model. If the stream ID is detected, it is always detected veridically - it is not possible to misidentify the stream ID of a word object. Also, the two-sided detection function is unusual in that it claims that the Stream ID can be detected better when the SNR is substantially negative, unlike almost every other detection function ever proposed. While the symmetry makes sense in terms of accounting for how a stream can be identified if it is the quieter of the two, this seems unnecessary because in this experiment, any time one stream is quieter than the other, the other is automatically louder. Why can't the stream ID attribute be detected according to a more conventional process?

Problems with predicting three- and four-speaker effects

The most serious problem was that the model did not scale to additional speakers. We made use of the data from Brungart, Simpson, Ericson, & Scott (2001) which is basically the same CRM paradigm except two or three masker speakers were included and the single-masker condition was represented by the same data as in Brungart (2001). We used the reported subset of the data in which the masker speakers were all of the same type - i.e. different talkers of different sex from the target talker, different talkers of the same sex as the target talker, and the same talker as the target talker. Unfortunately, in this paper the data is much less detailed; only the proportion of completely-correct responses (both color and digit are from the target message) is reported. Figure 6 shows this subset of the data from Brungart et al. (2001) for two-, three- and four-speakers (corresponding to one, two, and three maskers). As reported elsewhere in the literature, for more than two speakers, the proportion correct drops very quickly at low and negative SNR. In other words, adding a third speaker causes performance to fall off a cliff; a fourth speaker makes it only somewhat worse.

To extend the model to more than two speakers, we made two changes: (1) we modified the masking model to handle the case of more than two word objects masking each other. For each word object, an effective SNR was calculated by summing the total masking power in the other word objects so that having two maskers of the same

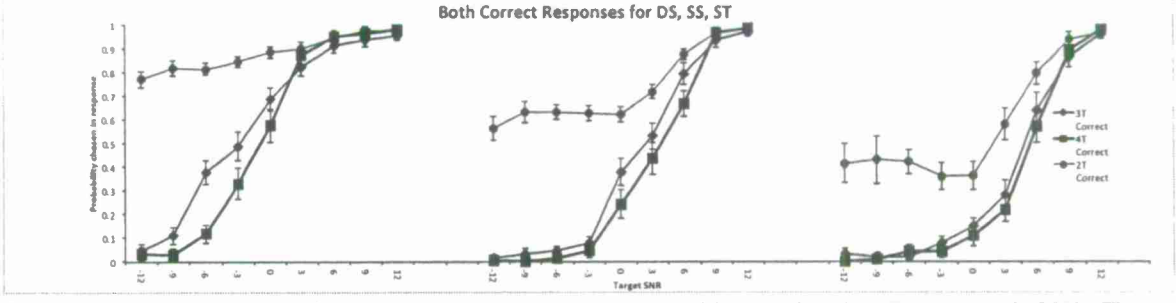


Figure 6. Observed proportions of both-correct responses for two, three, and four speakers from Brungart et al. (2001). The three speakers conditions are shown left-to-right in the three panels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). The highest performance is for two speakers; three and four speakers produce much lower, and similar, performance.

loudness decreased the SNR of the target by 3 dB. The content and stream ID detection functions used this effective SNR to determine whether content or stream ID was masked for a particular word object. The production rules were expanded to deal with multiple masker streams. This is relatively simple because the change is mostly in the rules for inferring target vs. masker content when some of the content or stream IDs are missing. For example, if the target stream could be identified from its callsign, then all other streams could be tagged as masker streams even if their callsigns were unheard. If all masker callsigns were heard and associated with a stream ID, than any "odd" stream had to be the target even it its callsign was unheard. However, notice that relative to the two-speaker case, such inferences are weaker. For example, suppose there are three speakers, and both the target callsign and one of the masker callsigns was unheard, then it is not possible to infer from the known masker stream which of these other two streams was the target. But if only two streams are involved, then the status of one can always be inferred from the status of the other.

This expanded model was run in the 3- and 4-speaker conditions and compared to the data reported in Brungart et al. (2001). Figure 7 shows what happened when the expanded production rule strategies were used with the same psychometric functions for content and stream ID detection for 3- and 4-speakers. The observed both-correct proportions are the black solid points and lines; the prediction values are the black open points with dotted lines. The other predicted values are coded the same as in Figure 4, but for brevity, only for predicted color responses. Since only the both-correct data was reported, the predictions for target responses, masker response, and neither responses cannot be compared to the data.

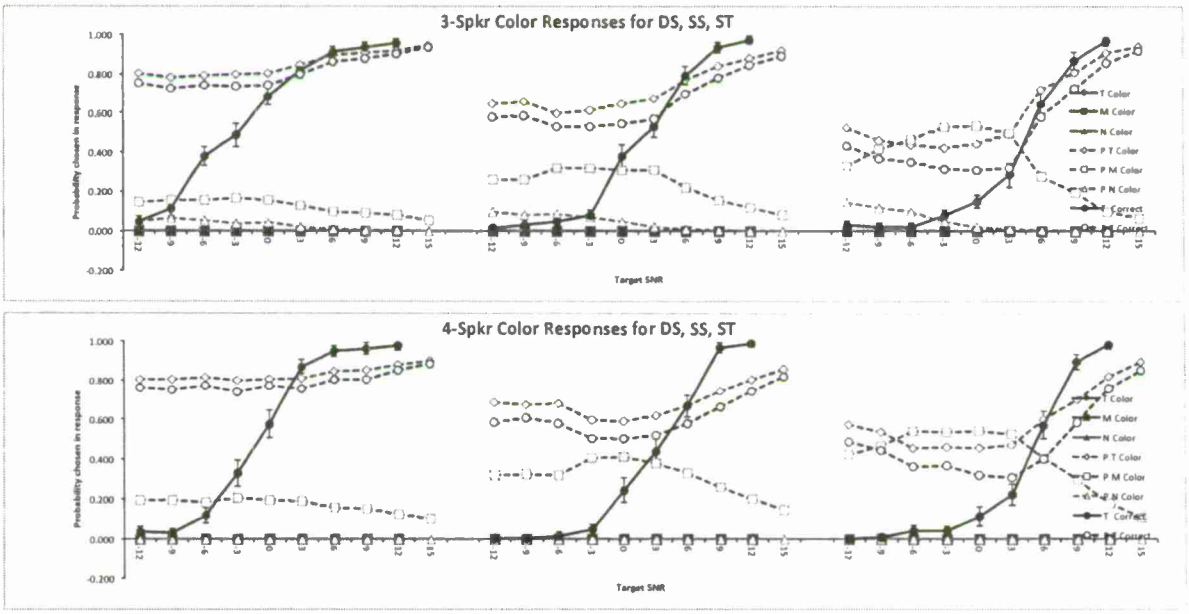


Figure 7. Observed (solid lines and points) and predicted (dotted lines and open points) probabilities for both-correct response choice in the Brungart et al. (2001) data. Results for three speakers are in the upper panel; four speakers are in the lower panel. The three speakers conditions are shown left-to-right in the three subpanels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). The blue, red, and green curves are for predicted target color, masker color, and neither, whose observed values were not reported. The black curves show the probability that both the color and digit are correct. Note the serious divergence of predicted and observed values, especially at negative SNR.

The problem is obvious - the observed both-correct proportion drops substantially at negative SNRs, but the predicted performance stays high, apparently due to the high sensitivity of stream ID detection at low SNR produced by the double-sided stream ID detection function. We investigated whether this was in fact the problem by fitting the three- and four-speaker data assuming a "normal" single-sided gaussian detection function for the stream ID. As shown in Figure 8, this produced a good fit to 3- and 4-speakers, but the same function applied to the two-speaker data was completely wrong. Keep in the mind that production rules are accounting for the strategy differences implied by the lesser ambiguity in the two-speaker case where inferences about missing information are stronger.

So the problem in moving from a single masker to multiple maskers was that the fundamental form of the stream ID masking function appears to be different - this is not a simple parameter value difference, but a fundamental difference in how the data could be modeled, even though we had the full power of EPIC's cognitive inferential machinery available to extract all of the information from the speech input. The difference between one masker and two maskers was not being captured by our modeling approach. This made us suspect that the underlying theoretical

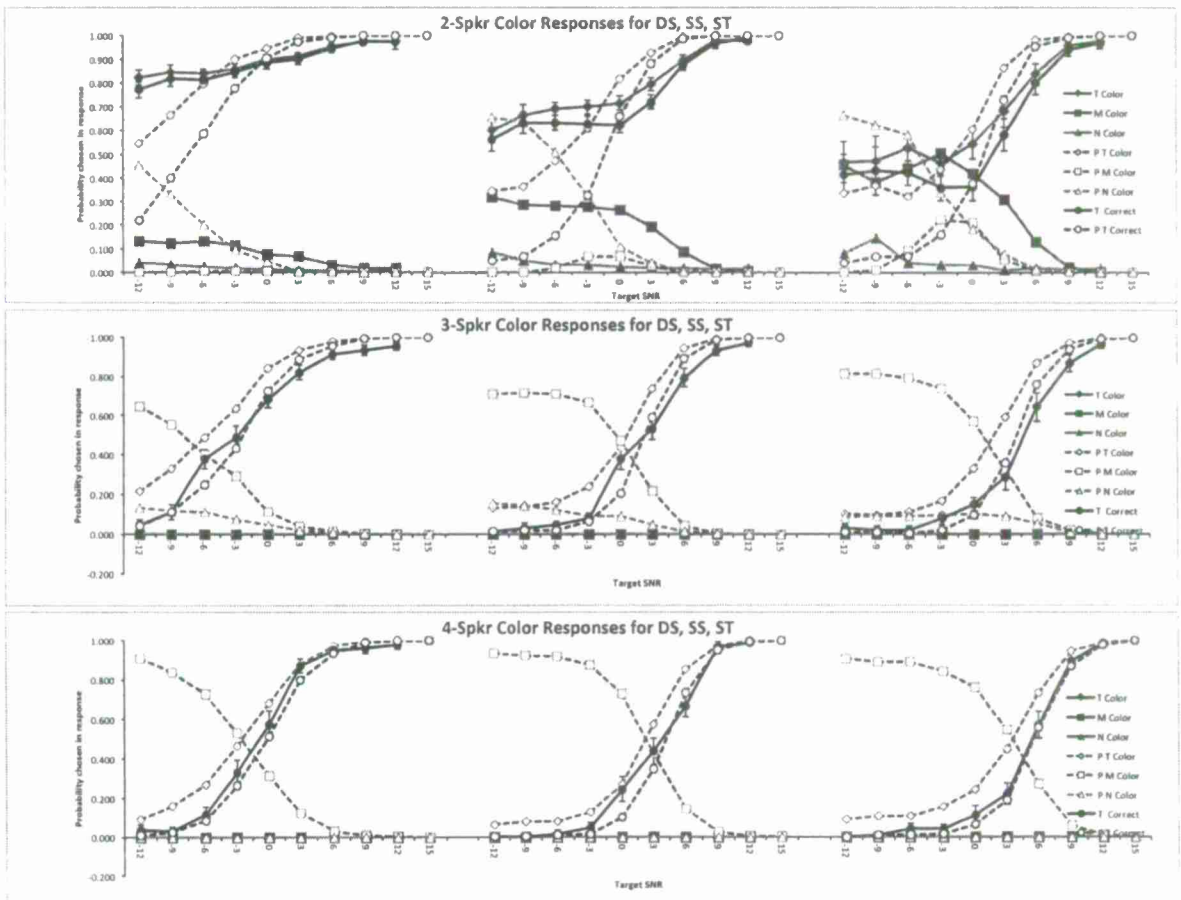


Figure 8. Observed (solid lines and points) and predicted (dotted lines and open points) response probabilities using simple gaussian detection functions for stream ID. Results for tow, three, and four speakers are shown in the top, middle and bottom panels. The three speakers conditions are shown left-to-right in the three subpanels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). The black curves show the probability that both the color and digit are correct. For two speakers, the separate color target, masker, and neither observed and predicted data are also shown, coded as in Figure 1. The both-correct (black) predictions are close to the data for three and four speakers, but seriously discrepant for two speakers.

structure of our model for the one-masker case was incorrect, even if it could be made to fit the data very well by an unusual detection function.

Reluctance to publish a model known to be defective

Because we doubted the validity of some of the basic assumptions, we decided not to attempt to publish our two-speaker model results, even though a model of this precision and explanatory power was unprecedented in this field. Rather we proceeded to rethink the fundamental structure of the model, in particularly, the decisions we had made in black-boxing the primitive perceptual processes.

Stream tracking concept

The problems with the black-boxing have already been alluded to above. Constructing a cognitive architecture and model for a complex psychological phenomenon has to meet practical constraints by picking what to black-box and what to actually represent. Our original concept for black-boxing the perceptual process was that each word object had two independent perceptual attributes: content and stream ID. Content was the recognized word, and the stream ID was a stand-in for whatever perceptual attributes allowed one to distinguish one speaker or message from another. We assumed that these attributes were either detected in veridical form, or not detected at all; that is, there is no possibility that a stream ID would be detected but it would the wrong stream ID.

The difficulty with our black-box assumption became apparent when we considered the case of the Same-Talker condition at SNR 0. The fits from our two-speaker model said that under these conditions the stream ID would be undetected most of the time. However, it seems intuitively obvious that a listener in this case would be able to tell quite easily that the two messages were from the same talker, which implies that the stream ID attributes were being detected, but simply could not be reliably assigned to appropriate words! This case suggests that the listener in this experiment always detects the stream-relevant acoustic properties of the words, but because these properties vary from one word to the next even for the same talkers, there is potential confusion as to which words come from which stream. However, there is some basis for distinguishing two utterances from the same talker - there is some tendency for messages from the same talker to differ in loudness and pitch contours.

We decided to apply a different black-box analysis of stream ID attributes. We assumed that the stream ID would be an inferred property that is based on acoustic attributes of the sound that are always detected. Two sound attributes that would be relevant in this type of study are the pitch and loudness of the sounds coming from the different speakers. For example, if the two speakers are different sexes, words from the female speaker are likely to have higher pitches than words from the male speaker. So the perceptual system could assign all higher-pitched words to one stream, and lower-pitched words to the other stream. Likewise, if one of the speakers is at a higher loudness than the other, the perceptual processor could assign all of the louder words to one stream, and the quieter words to the other. If the two speakers differ consistently in these attributes, the stream assignment of the words will be consistent and correct. However, if the two messages are similar in loudness or pitch, and are variable, then it is possible that at some point in the message, the stream assignment will "flip" or "switch" and the words will get assigned to the wrong streams.

Stream tracking model

We changed the auditory representation of the model. Instead of a stream ID attribute, each sound object carries its mean pitch and loudness as attributes. We assumed that these stream attributes are always detected. Stream

objects are no longer available to be "finessed" by being created under the control of the simulated environment. Instead, stream objects are only created by the auditory perceptual system whenever there is more than one sound object simultaneously present. Each stream object is a percept, and carries the loudness and pitch information accumulated for all the word objects that have been assigned to that stream. The first words in the two messages result in two sound objects that are simultaneously present; a stream percept is created for each word and initialized with the pitch and loudness information for its word. The next pair of words starts a process in which each word is assigned to its closest match (on some metric) with one of the stream percepts, which in turn is updated to reflect the loudness and pitch of its newly assigned word. Thus as the two messages are presented, each word gets assigned to a stream based on how well that word corresponds to the words that have already been assigned to the stream. If a word turns out to match the wrong stream better than the right stream, then an error can result.

This new model seems appealing, but it is rather more complex than our original model that was based on the detection of simple stream IDs; in particular, it must be determined what algorithms should be used to compare words to streams and update the stream information. Since both loudness and pitch are relevant, we needed to explore how both of these would play a role. In general, loudness differences are represented as usual in terms of differences in dB, and pitch differences are recoded into semitone differences. After considering a statistical approach popular in machine learning, we decided to first determine the feasibility of two very simple models. In a mean-based stream tracker, the stream percept is updated to hold the mean pitch and loudness of each word assigned to it. When two new words arrive, their distance from the two existing stream percepts are computed based on a weighted sum of the pitch and loudness, and each word is assigned to its closest stream. In a order-based stream tracker, we tried to capture a more qualitative stream assignment; the stream assignment algorithm orders the words from loudness&highest to softest&lowest, and likewise with the streams, and then assigns the words to streams in that order. For the mixed cases, such as loudest but not highest, the pitch and loudness are combined with a simple linear weighting and the objects are compared using that weighted sum.

Using corpus statistics to drive the new stream tracking model

A key idea of this new set of models is that stream tracking is driven by the properties of the speech signals, and these properties can be computed from the actual speech messages rather than estimated as parameters to fit the task performance data. For example, rather than estimating the probability that a female voice stream could be distinguished from a male voice stream, we could compute the fundamental pitch of male and female utterances in the CRM corpus, and supply these pitches to the model. If the female speaker always stays at a distinctly higher pitch compared to the male speaker, then the stream identification will be stable through the two utterances.

Thus to apply these stream-tracking models to the CRM task data, we needed to characterize the loudness and pitch of each utterance in the CRM corpus at an appropriate grain size. The brute-force way to do this would be to segment each utterance for the beginning and end of each word, and then compute loudness and fundamental pitch for each word, or possibly at a smaller frame size (e.g. 40 ms). Since this would be very laborious, we decided to try a quick and simple segmentation to find out whether the stream tracking models were likely to work. We took advantage of the fact that most speakers in the corpus appeared to speak the utterance in a steady rhythmic beat, where "go to" occupies a single "beat" as do the other words, as in [ready][callsign][goto][color][digit][now]. We divided each utterance into six equal-length segments corresponding to the six beats, and computed the average

loudness and fundamental pitch over each segment in each utterance. These corpus statistics were then supplied with each word (segment) that was "heard" by the simulated human's auditory processor.

Refinement and simplification of perceptual parameters

One additional new perceptual factor was included in the auditory perception module. Studies show that a difference in fundamental pitch improves the discrimination of simultaneous vowel sounds (surveyed by Darwin, 2008). As a simple way to incorporate this effect, we defined the effective SNR for two words to be the loudness SNR plus the weighted absolute difference in fundamental pitch in semitones. Thus if the target has a difference pitch than the masker, the content is more detectable. If there is more than one masker, the mean of their fundamental pitches is used to determine the pitch difference.

Because this model was driven by the corpus properties, there are relatively few free parameters that affect its fit to data. These are:

- The mean and standard deviation parameters for the ogival content detection functions for callsigns, color, and digit words, but unlike the previous models, these functions are assumed to be the same across the three speaker conditions (Different Sex, Same Sex, and Same Talker).
- The weight for loudness versus pitch differences in effective SNR.
- The weight for loudness versus pitch differences in determining the best assignment of word to stream.
- The specific rule for determining the best assignment.

Preliminary fits of new model

This model shows "signs of life" when compared to the two-speaker data as shown in Figure 9 below using an informal preliminary fit. While the fit is rather poor compared to the previous model in Figure 4, most of the effects are being determined by the acoustic properties of speech messages rather than a multitude of free parameters. The predictions for Different Sex (DS, leftmost panels) condition are in the right ballpark; the different pitches of male and female speakers make the content detection and stream tracking very robust. For the Same Talker condition (ST, rightmost panels), the predicted values are again in the right ballpark except at negative SNRs for Color. The Same Sex condition (SS, middle panels) are seriously discrepant at the middle SNRs, meaning that the model is not able to distinguish two speakers of the same sex from each other if the relative loudnesses are similar.

The failure of the stream tracking in the Same-Sex condition led us to look at what our computed corpus statistics were like. As expected, the female speakers are about an octave high in pitch than the male speakers. There is basically no overlap between genders, but within each gender, there is quite a bit of overlap, especially for females. In contrast, the loudness values overlap substantially. To some extent, this is due to how the corpus was prepared in order to allow the SNR to be manipulated; the average loudnesses of the utterances were supposed to be constant, but it tends to decrease by about 3 dB over the course of the utterance. It seems clear that using these statistics, the simple stream tracking algorithm could not tell speakers apart within gender. Development of a "smarter" stream tracking algorithm is part of the proposed work.

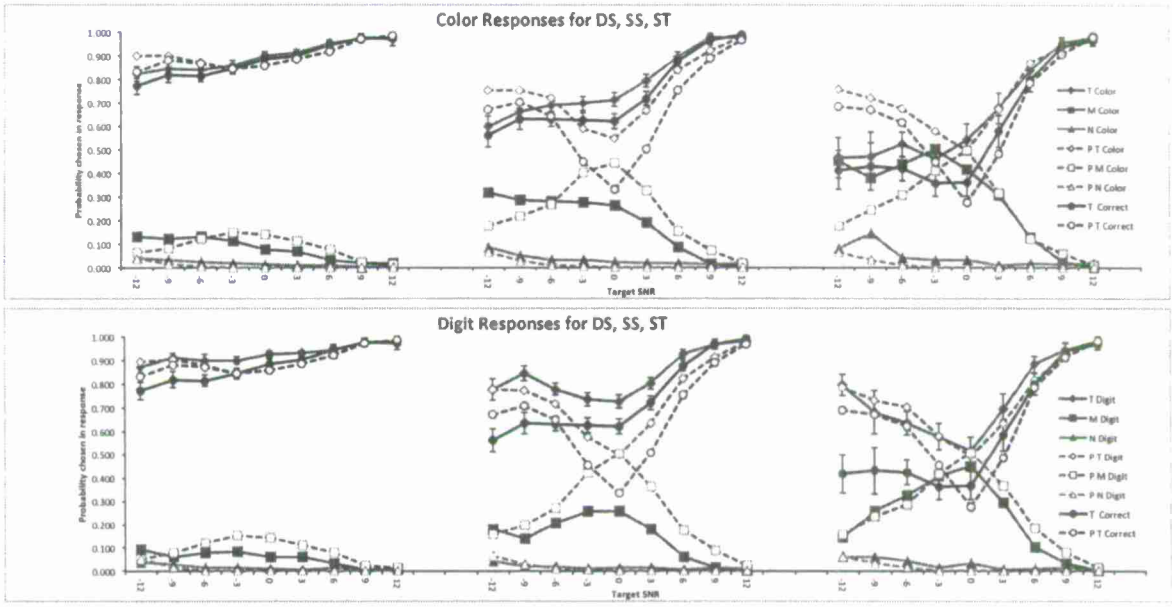


Figure 9. Preliminary fit using order-based stream tracker and six-beat segment analysis of the pitch and loudness of the CRM corpus. While there are problems with the fit, note that the model is driven by corpus statistics rather than a multitude of free parameters.

In the ball park for multiple maskers!

The impetus for abandoning our original simple representation of stream ID and going to a more complex, but intuitively compelling, notion of stream tracking was that our original model collapsed when we tried to scale it to multiple maskers in the three- and four-speaker case. The problem was that the stream ID detection functions had to be different forms in the single-masker versus multiple-masker conditions, a serious failure of basic parsimony and strong hint that something was wrong.

But as discussed above, the major effect of going from a single to multiple maskers is that performance drops substantially, either due to a perceptual problem, or the fact that the multiple maskers weaken the strategic inferences that can be made in performing the task. As mentioned above, the strategy differences are easily represented in the model, but don't explain the drop off in performance, so the problem must be perceptual, in that more simultaneous speakers makes each of them less perceptible. The problem with the earlier model was that rather than multiple speakers being less perceptible, the basic nature of the perception of the stream ID had to change - as reflected in the very different forms of the detection functions. But what if we represented the lower perceptibility of speech simply by less sensitive detection parameters for content in the presence of multiple speakers? Figure 10 shows a preliminary fit to the three-speaker data using the exact same model whose preliminary fit is shown above in Figure 9 with content detection functions that have the same ogival form, but are simply 3-8 dB less sensitive - a one-parameter change in the three content detection functions. This fit is also only preliminary, but unlike the complete failure of the previous model to apply to both single and multiple maskers (Figures 7 and 8 above), this model is producing "in the game" fits to single and multiple maskers, suggesting that the stream tracking approach is promising when combined with more complete models of how multiple maskers might affect content recognition. Refining the stream tracking model is thus our highest priority for continued work.

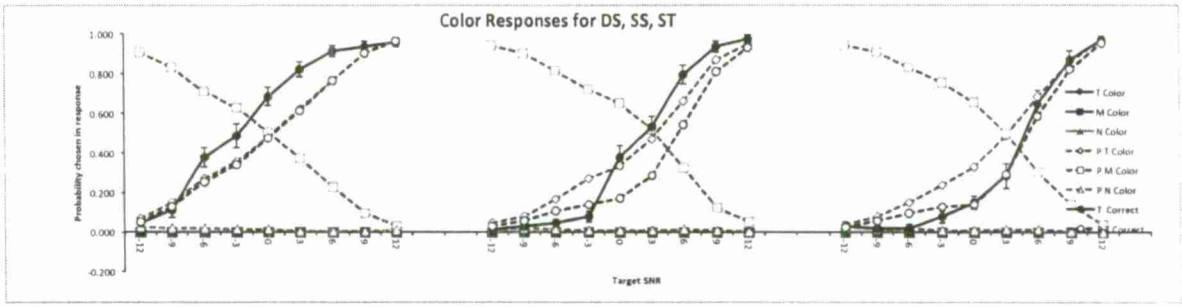


Figure 10. Preliminary fit of stream-tracking model to three-speaker data. The model is the same as in Figure 9, but with less sensitive content detection functions. Note that only the both-correct data (solid black lines and points) were reported and should be compared to the predicted values (dotted black lines, open points).

Determining the Plausibility and Requirements for a Stream Tracking Model

Despite its promise, before undertaking the complex work of developing a "smart" stream tracker, which might require working at the fine grain of the speech signal, we constructed and fit a mathematical simulation model to check that our stream tracking concept could indeed account for the data, and at the same time give us an advance reading on what characteristics the stream tracking needs to have in order to account for the observed performance data. The basic idea for this model was to assume that the stream tracker starts with the correct assignment of callsign words to streams, and then at the color word can "switch" to the wrong stream or "stay" at the correct stream, and then switch or stay again at the digit word. In the meantime, the content of the color and digit words is either recognized, or not, depending on a separate pair of probability functions. The model then uses the response rules from the complete EPIC model to choose its responses depending on the stream ID and content available for each color and digit word from each stream. Using (long-running) optimization routines in Matlab and Monte Carlo simulations, we estimated the probability functions for content recognition and stream-switching that fit the data, and we also compared hypotheses about task strategy as represented in the production rules.

This work was done in three steps with a model for each: In the first, we tested a model that assumed that word content was always recognized, so the only source of errors was stream assignment; the fit optimizer found the best fit given that only the stream "stay/switch" probability could account for the data. This model predicted zero Neither responses, a familiar failure mode of our models, and also required implausibly asymmetric stream "stay" probability functions (implausible because of the symmetry of the two-speaker task means that target and masker must symmetrically interfere with each other).

A second model included our usual assumptions that content was detected with a simple gaussian function, and the optimizer found the best-fitting parameters of this function for each speaker condition as well as the switch probabilities. In the remaining fits, we constrained the switch probability function to be symmetrical around 0 SNR, given the symmetry of the task. This model fit well except where the flatness and rise of the performance in some conditions at negative SNRs is present; as noted above, this is the most problematic aspect of the data.

The third model added an important innovation. Throughout this work, we have compared predictions made by task strategies that differed in how they choose "guesses" when the target stream content was missing - e.g. when the content of the color word from the target stream was not detected. In such cases of missing target content, how should a guess be chosen? One guessing strategy, called *avoid maskers*, optimizes against the likelihood of error by

rejecting words whose stream ID is that of the masker stream - better to make a guess from the other possibilities than to choose a color that we think is incorrect! But a different guessing strategy, called *use maskers*, seeks to improve the likelihood of a correct response by assuming that if the target content is missing, then maybe stream ID assignment was also incorrect, and so if other content is available, use it even if it is tagged as from the masker stream. This is a "use what you heard" strategy - something you heard is a better guess than a random one, even if it might be from the wrong source! These two strategies tend to produce different predictions of Masker and Neither responses, especially at negative SNRs.

The innovation in the third model is that rather than assume one or the other of these two guessing strategies, the model used a mixture of the two: if target digit content was missing, then with some probability, the avoid masker strategy was used, or the use masker strategy was used. The limitation of the mixture strategy to missing target digit content was another innovation; previously our guessing strategies had applied uniformly to both color and digit content. However, on average, the digit responses are more accurate than the color responses, even though there are more possible digits than colors. Some of our earliest work on the project showed that the digit words in the corpus were less co-articulated and possibly less synchronized than the color words, both of which would make the digits easier to recognize than the colors. In addition, note that digit content is more diagnostic of a correct response than color content. All of these differences justify treating missing target digit content differently from missing target color content in the guessing strategy.

Mixture models are used occasionally in production-system cognitive models to deal with the fact that in most psychological experiments, the subject's task strategy is neither assessed, trained, nor controlled, and so the data might reflect an unsystematic amalgam of different strategies. If only aggregated data is available, it is not possible to determine whether this strategy variation is within- or between-subject. Our assumption in this third model is that the variation is effectively within subject in that the mixture decision is made independently on each trial on which the guessing strategy is triggered. The fit optimizer adjusted the mixture probability along with the content detection and stream-switch probabilities.

The final result for the third model was a very close fit shown in Figures 11, 12, and 13 below which show the fit to the observed response probabilities and also the various probability functions used in the fit. Note that these are fits produced by the Matlab model rather than the full EPIC architecture model; the resulting parameters describe what the EPIC architecture version of the model needs to be able to do with the different messages in different speaker conditions. In Figure 11, the Different-Sex condition, the fit is extremely good. Content detection and stream assignment accuracy are high throughout because streams are well segregated due to the pitch difference. Content detection is only somewhat lower at the smallest SNR than highest. The estimated guessing strategy mixture probability varies somewhat at low SNR; at high SNR the estimate is unstable since a guess is rarely required.

In Figure 12, the Same-Sex condition, again the fit is extremely good. Content detection accuracy is fairly high throughout, but definitely lower than in the Different-Sex condition shown in Figure 11. Stream assignment accuracy is high, but drops somewhat when SNR is near zero, where the loudness difference is not reliable. The probability of staying at the previous (initially correct) stream is lower for colors than digits, meaning that if a switch happens, it is more likely to be at the color rather than the digit. The estimated guessing strategy mixture probability varies somewhat at low SNR and is similar to that obtained in Same-Sex fit. Again, at high SNR the estimate is unstable

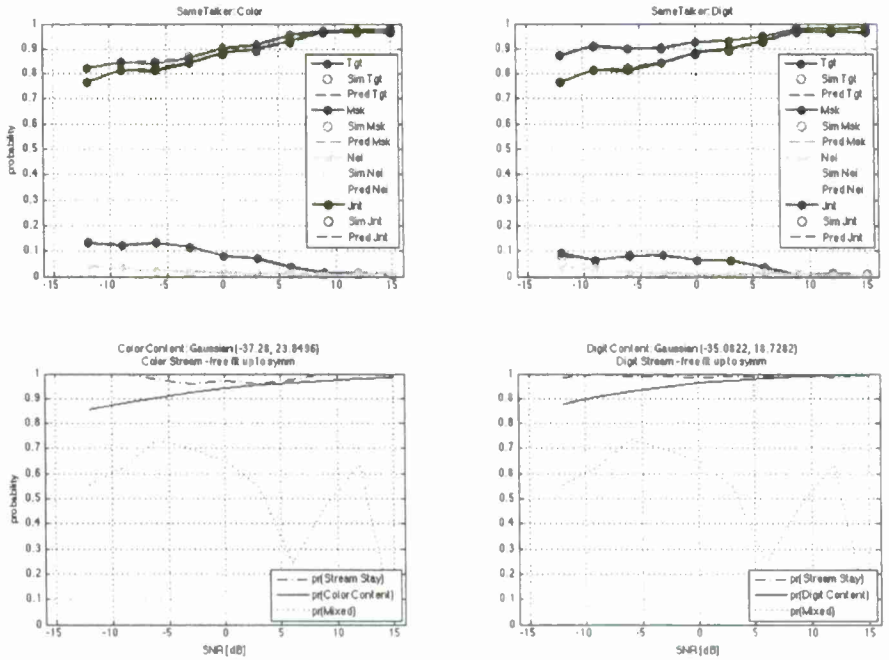


Figure 11. Results from data fitting using MatLab optimization on simplified model for the 2-speaker data in the Different-Sex speaker condition. The upper panels are observed (solid lines and points) and predicted (dashed lines & open points) response probability for target (blue) color (left) and digit (right), masker color and digit (red), joint target color and target digit correct (black), and neither target nor masker choice (green). Where a predicted curve cannot be seen, it is hidden under the observed data. The lower panels show the content detection functions for color (left) and digit (right) as the solid lines, the probability of staying on the current (initially correct) stream (large dashes), and the mixture probability for the guessing strategy choice (dotted).

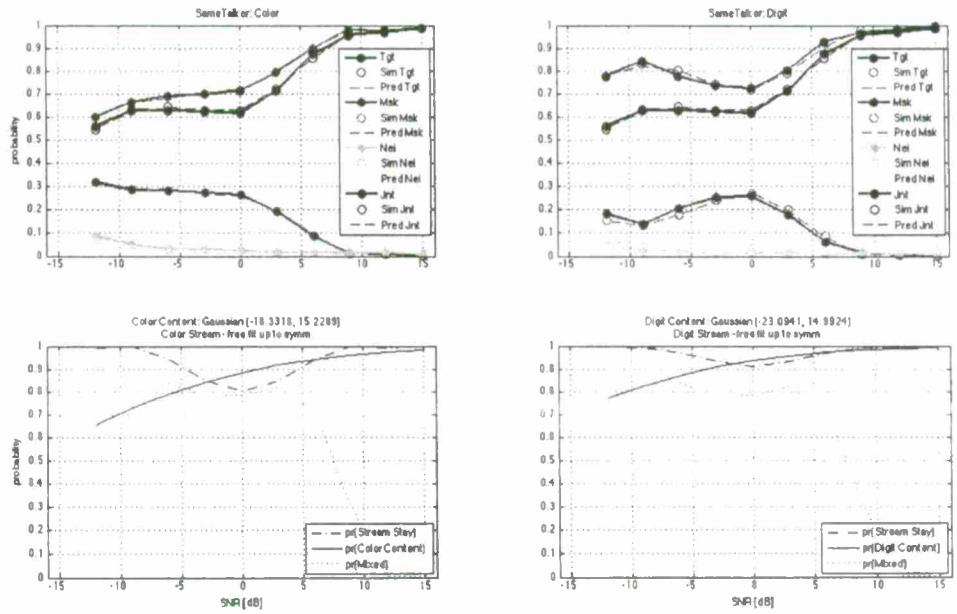


Figure 12. Results from data fitting using MatLab optimization on simplified model for the 2-speaker data in the Same-Sex speaker condition. Color codes are as described above in Figure 11.

since a guess is rarely required.

Finally, in Figure 13, in the Same-Talker condition, the fit is fairly good. Content detection is much lower than in the other conditions, and is much lower for colors than for digits, consistent with earlier analyses. The stream assignment accuracy drops substantially when SNR is near zero, where loudness difference is not reliable, much more so than for the Same-Sex condition. The probability of staying at the previous (initially correct) stream is also much lower for colors than digits, meaning that if a switch happens, it is more likely to be at the color rather than the digit. Estimated guessing strategy mixture probability varies more at low SNR than in the other conditions. Again, at high SNR the estimate is unstable since a guess is rarely required.

These results demonstrate that we should indeed be able to construct a model that accounts for the 2-speaker results, with their difficult and puzzling character, and also set some requirements for what the model will have to include. First, a mixture model for the guessing strategy is required that is triggered by missing target content in the most diagnostic case, the digit content. It is possible that a fixed mixture probability will account for the data reasonably well, which would mean that only a single parameter value for this factor will be needed. In addition, the architecture model should be able to fit the data with reasonable and well-behaved content detection functions. These functions will need to have different parameters for the different speaker conditions that reflect similarities in the speakers, but might be simplified as described below.

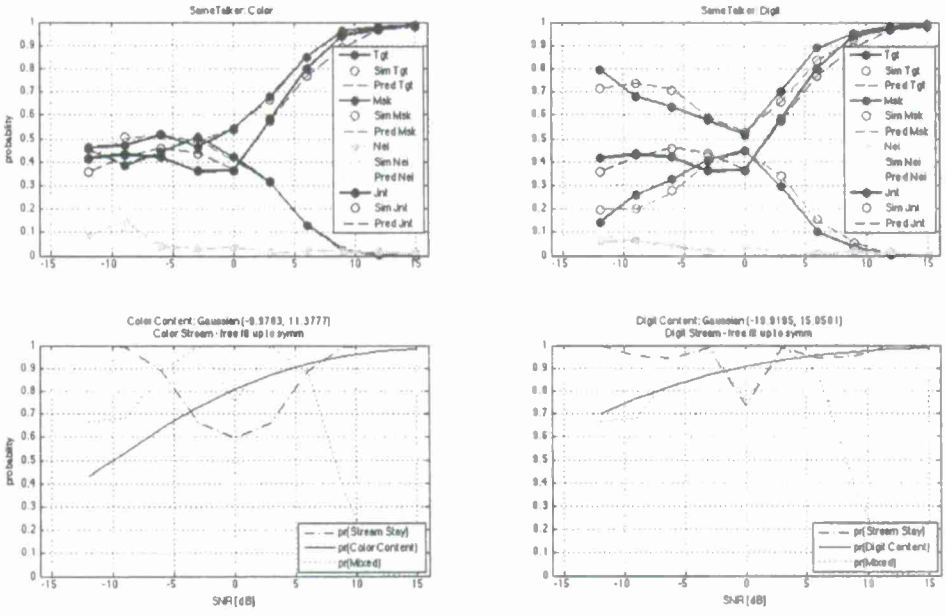


Figure 13. Results from data fitting using MatLab optimization on simplified model for the 2-speaker data in the Same-Talker speaker condition. Color codes are as described above in Figure 11.

Most importantly, the behavior of the stream switch/stay probability function needed to fit the data suggests two important requirements for the stream tracking mechanism:

- The stream assignment is generally reliable when the SNR is different from zero. This means that the effects of the different speaker conditions mainly involve whether the streams can be segregated when the loudnesses are similar. In the case of Different Sex, the consistent pitch difference suffices, and results in a very high stream assignment accuracy. In the case of Same Talker, the stream assignment accuracy is very poor, as expected, because all of the speech characteristics are similar. For the Same Sex condition, the stream assignment accuracy is intermediate between the other two conditions; if a stream attribute can be identified that produces assignment accuracy corresponding to the stream "stay" probability function in Figure 11, then this condition can be accounted for.
- The probability of a stream assignment switch between color and digit is relatively low, as shown by the high digit "stay" probability in the above figures. This means that whatever attributes are used to identify and assign the stream, they are stable going from color to digit, and more variable going from callsign to color.

Modeling simultaneous speaking and listening

Background

A problem emerged in the GLEAN modeling of AAW team tasks (Santoro, Kieras, & Pharmer, 2004; Kieras & Santoro, 2004). Each operator on the AAW team needs to speak messages on an external radio channel as part of their task, but at the same time, there might be incoming speech on the intercom to which they need to respond. At the time, the conflict was ignored in the modeling as it was not obvious how important this conflict would be, and, surprisingly, there exists a serious lack of research on the subject. In a seminal work, Broadbent (1952) reported a study of simultaneous speaking and listening in a simulated radio-communications task. His results suggest that under some conditions, the interference can be substantial.

Broadbent (1952) had subjects answer questions about which symbols appeared in columns of response sheet (Figure 14). A typical question was: *Hello, S-1. This is GDO. Is there an arrow in position 5? Over.* The task was first, to ignore the question if it was not addressed to the proper callsign (e.g. *S-1*). If the question was to answered, the answer had to include the correct sender callsign (e.g. *GDO*) as well as the correct yes/no response, and the question had to follow a radiotelephone protocol, for example: *Hello, GDO. This is S-1. Yes. Over.* A series of questions was presented that varied in whether they were to be ignored or answered and the sender callsign.

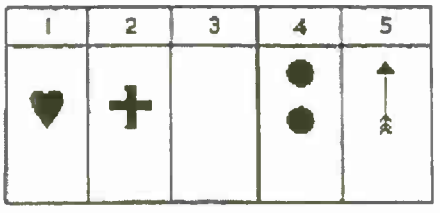


Figure 14. Example display used in Broadbent (1952).

Normally, there was a 4-sec interval between the end of a question and start of the next question (which was said to be ample time for the answer). Performance in this situation (isolated questions) functioned as a baseline control. Occasionally, the next question would appear immediately after the end of the previous question (overlapped), and the subject was instructed (and given practice) to answer the previous question while listening to the next question. Accuracy in this situation was decreased on both the previous question and the next question. In a final condition in the experiment, all of the questions were overlapped; performance was severely degraded.

A very compelling effect was that the accuracy depended heavily on the exact protocol used. There were three response formats: Long is like that in the examples above. In the Medium format, there were fewer "filler" words in both the question and answer, as in Q: *S-1 from GDO. Is there a heart in Position 1? Over.* A: *GDO from S-1. Yes. Over.* In the Short format, the questions were the same as the Medium format, but the answers were the shortest possible: A: *GDO. Yes*. Generally, the shorter the Q/A format, the more errors were produced.

There were a few follow-ups (e.g. Poulton, 1955), but almost immediately the task paradigm morphed into the "shadowing" task subsequently used to study attention effects, and no further studies were done with this clearly face-valid task. Oddly, similar to a point made much later by Poulton (1977) in a survey of the effects of noise on task performance, Broadbent argued that the observed effects had to be due to generalized attention phenomena rather than low-level masking effects that would seem to be obviously involved. As it happens, the empirical situation is unclear, which lead to a collaborative project with Nandini Iyer and Brian Simpson at the 711th Human Performance Wing, Human Effectiveness Directorate (711 HPW/RH), Warfighter Interface Division, Battlespace Acoustics Branch (RHCB), Wright-Patterson AFB.

A model for the Broadbent task

A preliminary version of a model for the Broadbent task was constructed, shown with the materials and detailed procedure in the new experiments. This model assumes that overt speech (produced by the subject) is treated as a speech stream perceived like external speech. Each word thus has a perceived content and stream identification, and this overt speech stream participates in masking during auditory perception like an external speech stream. Thus Broadbent's task is a form of 2-channel listening task, where one of the channels is generated by the subject who is also listening to the other channel. Thus if overt speech is produced at the same time as incoming externally-produced speech, words in the overt speech stream can mask the content or stream identification of the external speech. As an initial approximation, masking effects between overt and external speech were modeled using the same mechanisms as in the final version of the model for the Brungart (2001) data. Again for concreteness, Figure 15 shows a screenshot of the EPIC model of the Broadbent task, and Figure 16 shows a timeline of a specific overlapping question and answer with stream masking events, shown as lightning bolts, that cause errors on this trial. In addition, along the lines of the Kieras et al. (1999) models of verbal working memory, the information in auditory memory needed to answer the questions is subject to loss by decay over the few seconds required to answer the question. Thus even if there is no masking due to simultaneous speaking and listening, there will be some errors made.

The model strategy (see Figure 17 below) implemented two threads of processing: one to listen to external messages and identify the key content, which would reside in verbal working memory (the auditory memory); a second thread would decide whether to answer the question, and if so, would attempt to speak the answer words,

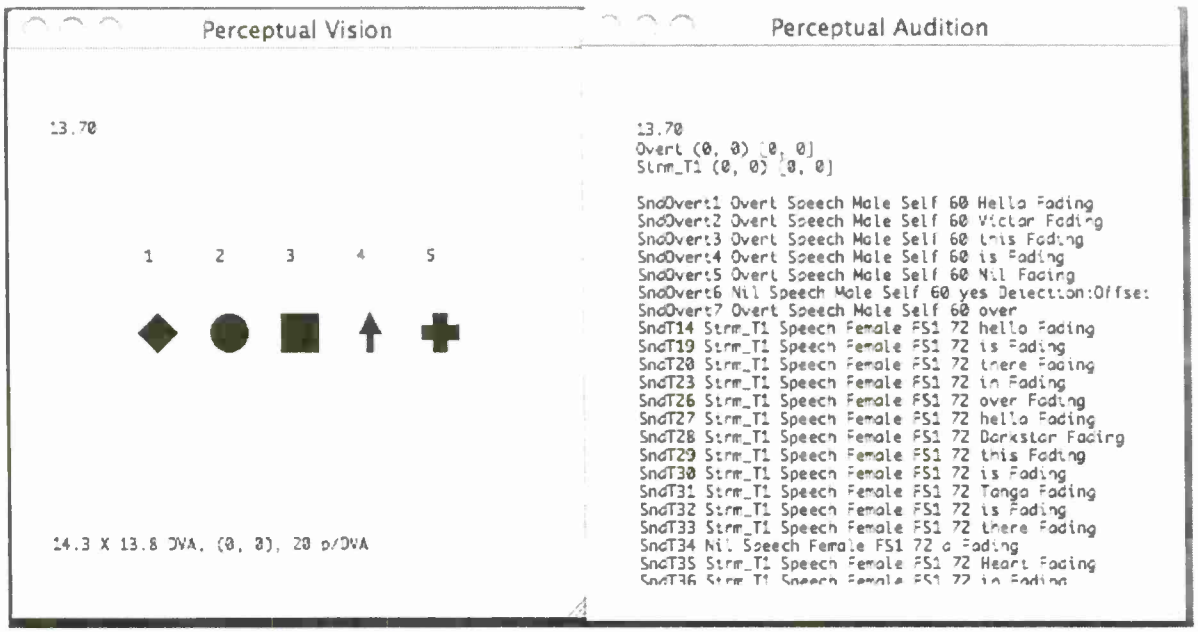


Figure 15. A screenshot of the EPIC model execution windows for the preliminary model of the Broadbent task. The visual display appears on the left; the current contents of the auditory perceptual store are shown on the right; auditory objects appear for both the external speaker asking the question and overt speech by the simulated subject providing an answer.

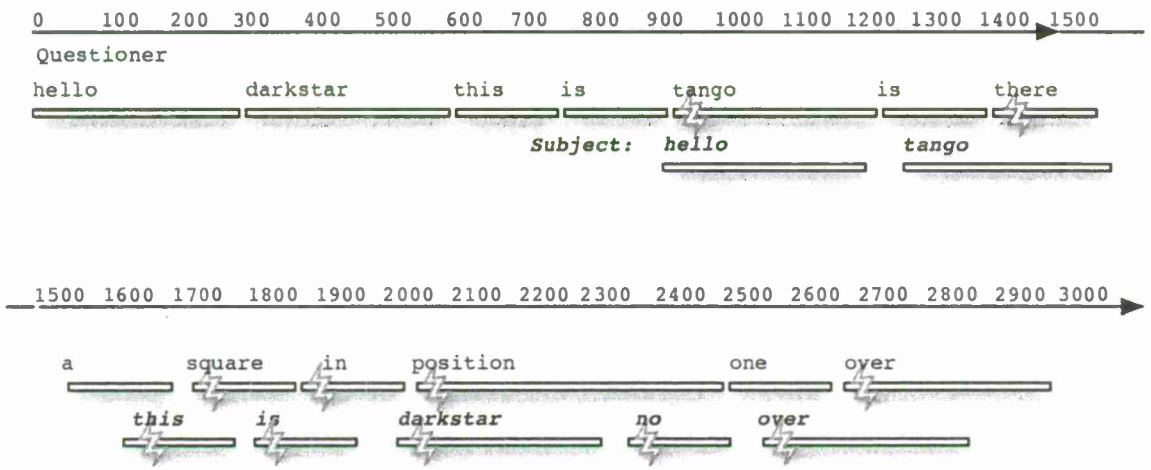


Figure 16. Timeline of an overlapping question and answer in the preliminary model for the Broadbent task. Cross-stream masking interference in this particular trial is symbolized by the lightning-bolt icons.

retrieving the content needed from auditory memory. If any content is missing, either because of masking or loss from auditory memory, the strategy is to guess the missing content.

The preliminary model makes a variety of errors like those reported in Broadbent, but it was not fitted to the data. It produces lower performance in overlapping questions than in isolated questions. The degradation in accuracy results from the masking of content words in the new question by the overt speech in a relatively haphazard way because the two speech streams are not synchronized, as in Brungart (2001). It should be the case that the exact

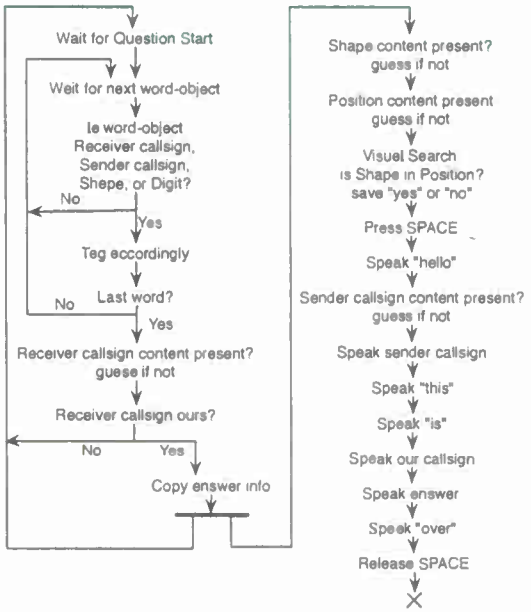


Figure 17. Task strategy used in the preliminary Broadbent task model.

word sequences used will produce different effects on performance due to different timing relationships in the masking process.

The task strategy required is surprisingly complex, not uncommon when such tasks are made fully explicit in a cognitive-architectural model (Kieras & Meyer, 2000), especially because since anything might get masked, the strategy has to provide for how to respond when words go missing. It is clear that there are a variety of strategy variations that will need to be explored in such tasks.

New results - failure of Broadbent to replicate

Two experiments were conducted by Iyer & Simpson as fairly direct attempts to replicate Broadbent's effects. To our surprise, the first experiment did not replicate Broadbent's effects! In fact, the performance was very high - there are very few errors made by the Wright-Patterson AFB subjects. This led us to reassess Broadbent's results, which following the custom of the time, were not reported as completely as one would currently expect. We estimated binomial confidence intervals for the reported proportions and sample sizes for the effects in Broadbent. The left-hand panel Figure 18 shows these; it is not necessary to use space in this report to explain the effects shown; suffice it to say that most of the effects at low error rates are not statistically reliable, so we should not be surprised if they did not replicate. Figure 18 shows the corresponding results from the first Iyer & Simpson experiment (unfortunately also without confidence intervals at this point, but very few differences are reliable). The error rates are much lower, and the trends are quite different. It is axiomatic that errors are very hard to study experimentally; the better subjects are trained, and the cleaner the experimental procedure, the fewer errors are made, and even if the samples are big enough to detect effects, at low rates, the errors will be of little practical importance.

The second experiment was run to check out a possible methodological problem: In the first experiment, the subject's own voice was not fed through their headphones - there was no "side tone". This could have eliminated any purely acoustic inference of their self-speech with what they were hearing. In the second experiment, the side tone

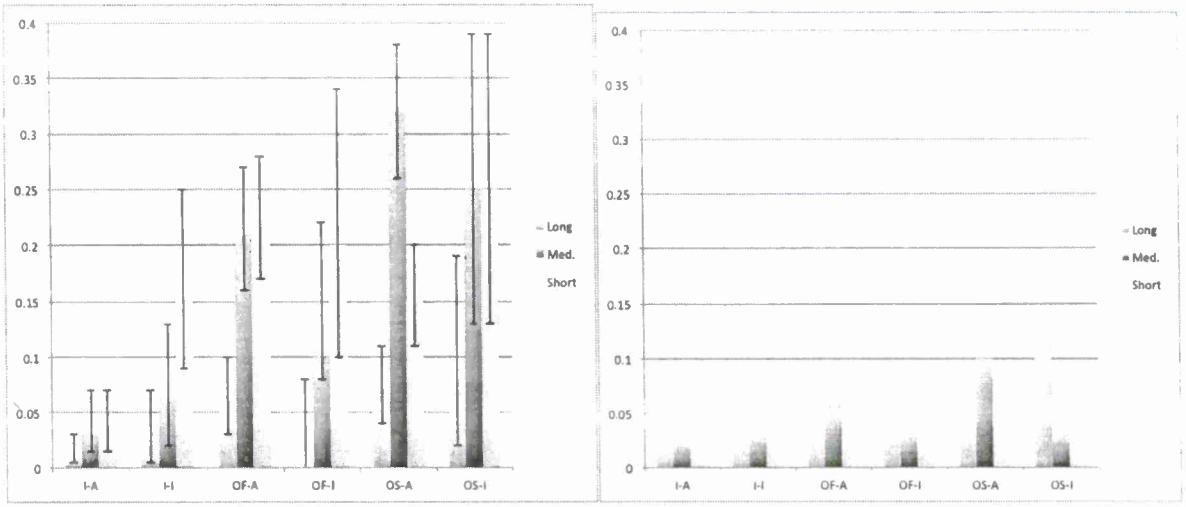


Figure 18. Error rates in questions answered under different conditions in Broadbent(1952) with estimated confidence intervals (left panel), and in Experiment 1 conducted by Iyer & Simpson at Wright-Patterson AFB (right panel).

was present and set at a fairly high level. The result was basically unchanged - again there are very few errors and no clear trends.

This result is counter-intuitive - surely speaking while somebody else is talking to you should result in rather poor comprehension of what they were saying (in addition to social effects)! Were Broadbent's results a fluke? Or do they depend on very special conditions? For example, since Iyer & Simpson's subjects are very well practiced, it is possible that they are able to develop strategies for doing the task that prevent interference - such as timing their speech to leave gaps for hearing the incoming message. Or perhaps they can somehow factor out their own speech signal from what they are hearing.

At this point, further work on modeling performance in simultaneous speech and listening was suspended awaiting an empirical resolution to this problem. Work has continued under collaborative project funding to conduct new experiments to determine the conditions under which interfere occurs during simultaneous speaking and listening.

Goal 4. The models and architecture will be developed and tested in complex military-like tasks, e.g. based on Combat Information Center (CIC) Anti-Air Warfare (AAW) tasks, based on earlier work with the Multi-Modal Watch Station (MMWS) project.

Introduction

Opportunities to interact with military researchers are both rare and valuable because there needs to be a favorable confluence of good data, relevant task, and a feasible modeling problem that addresses an important issue, either theoretical, practical, or both. During this project, two significant activities occurred.

Complex CIC task with spatialized speech and auditory signals

NRL, Derek Brock's group. This line of work is focused on the type of task situation studied in the MultiModal Watch Station (MMWS) project, that involved Anti-Air Warfare CIC tasks that Kieras has had considerable modeling experience in. A specific example of how such work might progress is as follows: Computer-serialized speech input appears to produce superior performance compared to concurrent speech input in Brock et al. (2008). But how would serialized input be integrated into a complex task? Would the serialization facility help or hinder the user in time-critical situations? For example, suppose a "vampire" call (incoming missile warning) comes in on one channel at the same time as routine reports are being received on other channels. With concurrent presentation, a skilled watch stander could pick up the "vampire" call and switch attention to that channel, but would an automatic serialization system simply delay that channel? How could this be made to work? Different ideas could be tested by setting up EPIC simulations with different device designs. Because the real-world requirements for device designs can change quickly, our goal would be not so much to arrive at a specific good design, but rather to test whether the architecture provides good support for predictive models of device designs involving speech interaction.

Brock's group was prototyping complex workstations using multiple visual and auditory displays to present a radar-type display, chat windows, and spatialized and serialized speech channels that can also appear in text windows. As of summer of 2012, the prototype was awaiting additional scenarios before data collection could begin. While this domain is very complex, it is also clearly the kind of domain that should lie within EPIC's competence to represent, especially with its newer auditory and speech capabilities. In terms of future work, when the task specifications and data become available, if there are important effects, the NRL task would be a good candidate for a modeling effort.

Auditory-visual integration in new periscope display systems

NSMRL, Michael Qin's group. New digital imaging technology makes it possible to revolutionize how periscope observation is done - in particular, rather than a single observer physically walking the periscope around, it is now possible to collect a high-resolution 360° of image from a camera at the periscope head and feed it to multiple displays. But it is not obvious how to map a 360° view to one or two conventional flat screen displays, which is required by the space available on existing submarines. Furthermore, in some critical tasks, such as checking for nearby vessels before surfacing, sound information is available that could help guide the initial observation. Thus an additional question is whether and how this information could be presented as spatialized sound to improve the periscope task performance. The more general problem is how to answer such design questions by the use of human performance modeling in addition to prototyping and testing candidate designs with human users.

Qin is developing a laboratory experiment to collect some data on basic design alternatives for new periscope displays using a lab-scale version of the surfacing task. Kieras developed a first approximation EPIC model for this task, including the use of spatialized sound, which has to be recoded from a 360° auditory space to the rather small visual space on the display. Figure 19 presents the model display and explains the procedure for this laboratory simplification of the surfacing task. Figure 20 shows predictions for the effect of sound on reaction time as a function of the number of contacts presented on the display. The model predicted the intuitive result that the auditory signal would greatly speed up locating the target object on the display, but also estimated the *size* of the effect, and clarified that it would depend on the difficulty of visually classification and the difficulty of the sound encoding.

Kieras visited Qin's lab during the summer of 2012 to hand off the first-pass model and prepare for additional collaboration. Continuing this line of work would be advantageous scientifically because it would help develop and validate some applied models of visual search and spatialized audio aiding. An additional practical result of this work will be to make the EPIC project results and the simulation software more available to Navy R&D.

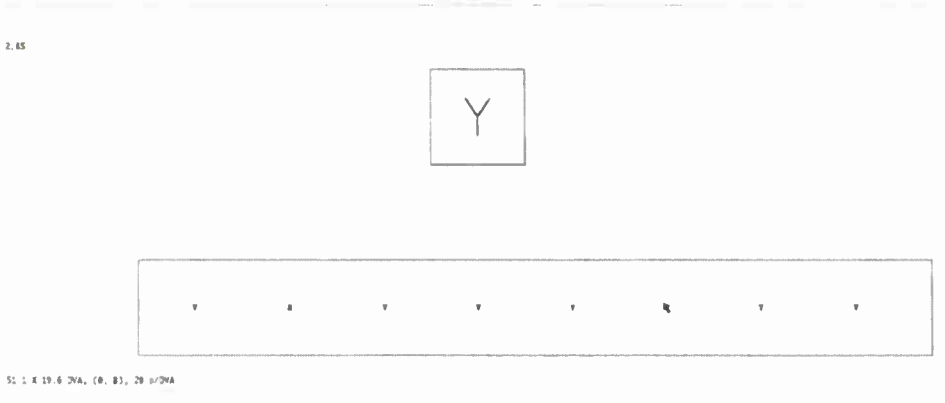


Figure 19. Display produced by the first-approximation model for the audio-visual integration experiment for 360° periscope-like displays. The long rectangle represents the panoramic 360° presentation with 8 contacts, only one of which is the target X shape. Selecting a particular contact presents a blown-up view in the top window. The trial continues until the subject has selected the target. An array of 8 sound sources surrounding the subject can cue the location of the target.

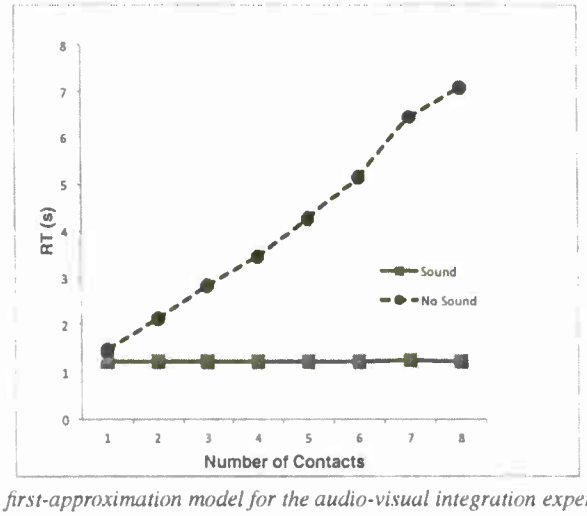


Figure 19. Predictions from first-approximation model for the audio-visual integration experiment for 360° periscope-like displays. The sound cue produces a very fast predicted RT that does not depend on the number of contacts, while no sound cue requires visual search that is linear with the number of contacts because of their small visual size.

References

Ballas, J., Brock, D. Stroup, J. Kieras, D. and Meyer, D. (1999) Cueing of Display Objects by 3-D Audio to Reduce Automation Deficit. In *Proceedings of the Fourth Annual Symposium and Exhibition on Situational Awareness in the Tactical Air Environment*. The Naval Air Warfare Center, Patuxent River, MD, June 8-9, 1999, pp 100-110.

Ballas J. A., Kieras, D. E. & Meyer, D. E. (1996). Computational modeling of multimodal I/O in simulated cockpits. In S. P. Frysinger & G. Kramer (Eds.) In *Proceedings of the Third International Conference on Auditory Display*. Xerox Palo Alto Research Center, Palo Alto, CA, Nov. 4-6, 1996, pp 135-136.

Broadbent, D.E. (1952). Speaking and listening simultaneously. *Journal of Experimental Psychology*, 43, 267-273.

Brock, D., Ballas, J.A., Stroup, J.L., & McClimens, B. (2004). The design of mixed-use, virtual auditory displays: Recent findings with a dual-task paradigm. *Proceedings of the 10th International Conference on Auditory Display*, Sydney, Australia.

Brock, D., McClimens, B., Trafton, J.G., McCurry, M., & Perzanowski, D. (2008). Evaluating listener's attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech. *Proceedings of the 14th International Conference on Auditory Display*, Paris, France.

Brungart, D.S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* **109** (3), 1101-1109.

Brungart, D.S., Simpson, B.D., Ericson, M.A., & Scott, K.R. (2001) Informational and energetic masking effects in the perception of multiple simultaneous speakers. *J. Acoust. Soc. Am.* 110 (3), 1101-1109.

Darwin, C.J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society: Biological Sciences*. **363**, 1011-1021.

Hornof, A. J., & Zhang, Y. (2010). Task-constrained interleaving of perceptual and motor processes in a time-critical dual task as revealed through eye tracking. Proceedings of ICCM 2010: The 10th International Conference on Cognitive Modeling, Philadelphia, Pennsylvania, August 5-8, 97-102.

Hornof, A. J., Zhang, Y., Halverson, T. (2010). Knowing where and when to look in a time-critical multimodal dual task. Proceedings of ACM CHI 2010: Conference on Human Factors in Computing Systems, New York: ACM, 2103-2112.

Kieras, D.E. (2007). The control of cognition. In W. Gray (Ed.), *Integrated models of cognitive systems*. (pp. 327 - 355). Oxford University Press.

Kieras, D. (in press). A summary of the EPIC Cognitive architecture. In S. Chipman (Ed.) *Handbook of Cognitive Science*, Oxford.

Kieras, D.E., Ballas, J.A., & Meyer, D.E. (2001). Computational models for the effects of localized sound cuing in a complex dual task. (EPIC Tech. Rep. No. 13, TR-01/ONR-EPIC-13). Ann Arbor, University of Michigan, Electrical Engineering and Computer Science Department. January 31, 2001.

Kieras, D., & Knudsen, K. (2006). Comprehensive Computational GOMS Modeling with GLEAN. In *Proceedings of BRIMS 2006*, Baltimore, May 16-18, 2006.

Kieras, D.E., & Meyer, D.E. (1995). Predicting performance in dual-task tracking and decision making with EPIC computational models. *Proceedings of the First International Symposium on Command and Control Research and Technology*, National Defense University, Washington, D.C., June 19-22. 314-325.

Kieras, D. E., & Meyer, D. E. (2000). The role of cognitive task analysis in the application of predictive models of human performance. In J. M. C. Schraagen, S. E. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum, 2000. 237-260.

Kieras. D.E. & Santoro, T.P. (2004). Computational GOMS Modeling of a Complex Team Task: Lessons Learned. In *Proceedings of CHI 2004: Human Factors in Computing Systems*. New York: ACM, Inc.

Kieras, D.E., Meyer, D.E., Mueller, S., & Seymour, T. (1999). Insights into working memory from the perspective of the EPIC architecture for modeling skilled perceptual-motor and cognitive human performance. In A. Miyake and P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York: Cambridge University Press. 183-223.

Kieras, D.E., Wood, S.D., Abotel, K., & Hornof, A. (1995). GLEAN: A Computer-Based Tool for Rapid GOMS Model Usability Evaluation of User Interface Designs. In *UIST'95: Proceedings of the ACM Symposium on User Interface Software and Technology*, New York: Association for Computing Machinery, pp. 91-100.

Kieras, D.E., Wood, S.D., & Meyer, D.E. (1997). Predictive engineering models based on the EPIC architecture for a multimodal high-performance human-computer interaction task. *ACM Transactions on Computer-Human Interaction*. 4, 230-275.

Meyer, D. E., & Kieras, D. E. (1997a). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*. **104**, 3-65.

Meyer, D. E., & Kieras, D. E. (1997b). A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of Psychological Refractory-Period Phenomena. *Psychological Review*. 104, 749-791.

Meyer, D. E., & Kieras, D. E. (1999). Precis to a practical unified theory of cognition and action: Some lessons from computational modeling of human multiple-task performance. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII*. (pp. 15-88) Cambridge, MA: M.I.T. Press.

Mueller, S. (2002). The roles of cognitive architecture and recall strategies in performance of the immediate serial recall task. Doctoral dissertation. Ann Arbor, Michigan: University of Michigan.

Poulton, E.C. (1955). Simultaneous and alternate listening and speaking. *Journal of the Acoustical Society of America*. 27, 1204-1207.

Poulton, E.C. (1977). Continuous intense noise masks auditory feedback and inner speech. *Psychological Bulletin*, 84, 977-1001.

Santoro, T.P., Kieras, D.E., Pharmer, J.A. (2004). Verification and validation of latency and workload predictions for a team of humans by a team of computational models. *U.S. Navy Journal of Underwater Acoustics (JUA(USN)) Special Issue on Modeling and Simulation*. 54. 281-304.

Products

Graduate Student Statistics

Consistent with the project budget, there were no graduate students supported by this project.

Journal Articles

Kieras, D. (2011). The persistent visual store as the locus of fixation memory in visual search tasks. Cognitive Systems Research, 12, 102-112.

Book Chapters

Kieras, D.E. (2012). Model-based evaluation. In J. Jacko (Ed.), *The Human-Computer Interaction Handbook* (3rd Ed). New York: Taylor & Francis. pp. 1299-1318.

Kieras, D.E. (in press). A summary of the EPIC cognitive architecture. In S. Chipman (Ed.), *The Oxford Handbook of Cognitive Science*. New York: Oxford University Press.

Kieras, D.E., & Butler, K.A. (in press). Task analysis and the design of functionality. In H. Topi (Ed.) *Computer Science Handbook, Vol. 2: Information Systems and Information Technology*, CRC Press.

Invited Presentations

Kieras, D. Human Performance Modeling with a Cognitive Architecture: Crystalizing the Science for Application. Keynote Talk presented at the 55th Annual Meeting of the Human Factors and Ergonomics Society, Sept. 20, 2011.

Kieras, D. Don't Pay Attention: Modeling Perceptual Selection with a Cognitive Architecture. Invited symposium presented at Georgia Institute of Technology, Nov. 9, 2011.

Kieras, D. Human Performance Modeling with Cognitive Architectures. Invited symposium presented at Michigan Technological University, April. 11, 2012.

Wakefield, G. & Kieras, D Auditory modeling in EPIC. STTR Kickoff Meeting, Wright-Patterson AFB, Feb. 3, 2012.

Awards/Honors

David E. Kieras: *Jack A. Kraft Innovator Award*, presented by the Human Factors and Ergonomics Society, September 2010, "in recognition of significant efforts to diversify and extend the application of human factors principles to new areas of endeavor."

David E. Kieras: *Election to the SIGCHI Academy* by the ACM Special Interest Group for Computer-Human Interaction, April 2010, "for leadership in the profession of Computer-Human Interaction."

Final Report Distribution List

Paul Bello

Office of Naval Research

875 N. Randolph St.

Arlington, VA 22203-1995

Office of Naval Rescarch

Regional Office - Chicago-N62880

230 South Dearborn, Room 380

Chicago IL 60604-1595

Defense Technical Information Center

8725 John J. Kingman Road STE 0944

Fort Belvoir, VA 22060-6218

Naval Research Laboratory

Attn: Code 5596

4555 Overlook Avenue SW

Washington, DC 20375-5320